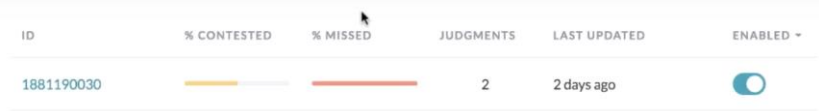
Project Proposal

Tamer Abdelaty Ahmed

Data Labeling Approach

Project Overview and Goal **What is the industry problem you are trying to solve? Why use ML in solving this task?**	Industry – Medical Problem – Pneumonia detection at scale requires great Human resources (lots of doctors), long time which increases pre-treatment cost, and diverts vital resources from the process of curing. Goal: 1. As a product owner is to build a product that helps doctors quickly identify cases of pneumonia in children. 2. build a classification system that • Can help flag serious cases • Quickly identify healthy cases • And, generally, act as a diagnostic aid for doctors Strengths: Fast and accurate detection of Pneumonia by scanning chest/lung X-rays using Machine Learning (ML) could address the above problems. The use of ML will increase the speed of classification of cases among those with pneumonia, pneumonia-free, and unclear requiring further investigation. Thus, accelerates the classification process, saves time, reduces cost, and potentially increases accuracy by eliminating human error. Weaknesses: capture uncertainty (Unknown)in labels is not always 100% correct.
Choice of Data Labels **What labels did you decide to add to your data? And why did you decide on these labels' vs any other option?**	The labels we use to classify our data is based on a binary Yes/No (Pneumonia/Pneumonia-Free) and an additional Inconclusive. We used the "Image Categorization method". An alternative method would have been to use a measure of degree of confidence (using the Search Relevance method – Rate the Relevance of a Search Result) with the following scale 1- Yes (Pneumonia) A: High (confident Pneumonia) B: low (Not Confident Pneumonia) C: Unknown 2. No (Pneumonia-Free (Healthy)) A- high (Confident Pneumonia-Free (Healthy)) B- low (Not Confident Pneumonia-Free (Healthy))

Test Questions & Quality Assurance

Number of Test Questions Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?	8 questions representing approximately 8% of the data set. I tried to represent all cases as I had the option to select representative cases for a balanced training. The mistake of using a bigger sample could lead to a biased model if one case was overrepresented due to the unequal distribution of cases within my sample.
Improving a Test Question Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question?	 1. The following are assumptions based on 100% of annotators missed. I did not run my model. 2. Doublecheck if the “pictures-samples” correspond to the desired outcome and the picture-samples are not corrupted. 3. Doublecheck if the correct answer corresponds to the picture and there has not been a human error during the assignment of meaning to picture. 4. Doublecheck that enough representations of each case are in place. 5.Doublecheck if corresponding explanations are given when both a Yes or a No selection was made to give context and increase confidence of the non-expert annotator.
Contributor Satisfaction Say you’ve run a test launch and gotten back results from your annotators; the instructions and test questions are rated below 3.5, what areas of your Instruction document would you try to improve (Examples, Test Questions, etc.)	 1. I would try to improve Examples, Questions, Instruction Document and try to show how to recognize unclear Examples. 2. I might start by providing more examples if the job is complicated or add more samples that require more attention to detail.

Limitations & Improvements

Data Source Consider the size and source of your data; what biases are built into the data and how might the data be improved?	The Dataset might suffer from the following any of these biases: 1.Too many samples of Pneumonia could lead the system to see "everything" as pneumonia. 2.Too many samples of non-pneumonia could lead the system to see "everything" as healthy. 3.Too many samples of "unclears" could lead the system to be indecisive. Improvements can be made: 1. by making sure the training sample remains fair in size and equally distributed among all the classes. 2.The size of the data set might be non-representative because the sample itself is too small or it is biased – which leads to the problem mentioned above. 3. Improvements can observe by increasing the size and making sure enough representations of all classes are included. 4. The source of the Dataset is contaminated/corrupted hence the data themselves will not be clean and thus improvements would require further human intervention to clean the data before classifying them.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	1. By observing the evolution of my data as if they are dynamic and their status is evolving / changing. 2. So I should also make sure the labelling is evolving/changing and thus my classification remains relevant or needs to be re-done. 3. More labelling will require possibly greater in number questions and most likely more questions of greater diversity.